Preschool/Kindergarten teachers' conceptions of standardised testing

Frans, Niek; Post, W. J.; Oenema-Mostert, C. E.; Minnaert, A. E. M. G.

Published in:
Assessment in Education: Principles, Policy & Practice

DOI:
10.1080/0969594X.2019.1688763

Publication date:
2020

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Preschool/Kindergarten teachers' conceptions of standardised testing

Niek Frans, W. J. Post, C. E. Oenema-Mostert & A. E. M. G. Minnaert

To cite this article: Niek Frans, W. J. Post, C. E. Oenema-Mostert & A. E. M. G. Minnaert (2020) Preschool/Kindergarten teachers' conceptions of standardised testing, Assessment in Education: Principles, Policy & Practice, 27:1, 87-108, DOI: 10.1080/0969594X.2019.1688763

To link to this article: https://doi.org/10.1080/0969594X.2019.1688763

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
2020, VOL. 27, NO. 1, 87–108
https://doi.org/10.1080/0969594X.2019.1688763

Routledge
Taylor & Francis Group

Preschool/Kindergarten teachers' conceptions of standardised testing

Niek Frans[a], W. J. Post[a], C. E. Oenema-Mostert[a,b] and A. E. M. G. Minnaert[a]

[a]Special Needs Education and Youth Care, University of Groningen, Groningen, Netherlands; [b]Academy of Teacher Education, Stenden University of Applied Sciences, Leeuwarden, Netherlands

ABSTRACT

Standardised tests play an important role in early childhood (EC) education in many countries. Although teachers' conceptions largely determine whether and how these instruments are used, research on this topic is scarce. As a result, factors that influence conceptions of standardised testing have remained largely unexplored. To examine teachers' conceptions of standardised testing and aspects that may influence these conceptions, Brown's CoA-III-A questionnaire was distributed to 97 EC educators. Based on their responses, a selection of six preschool/kindergarten teachers participated in a series of semi-structured interviews. Analyses of the questionnaire and the interviews indicated that the teachers did not see these tests solely as instruments for accountability or improvement. While some perceived the test as pleasant confirmation, others perceived the results as negative opposition to their own observations. The teachers' conceptions were influenced by classroom population, management team, and the ascribed purpose of the test.

ARTICLE HISTORY
Received 17 October 2018
Accepted 22 October 2019

KEYWORDS
Standardised testing; early childhood; teacher conceptions

Introduction

Given the key role that teachers play in educational assessment, their conceptions of the purposes of assessment influence how teachers filter assessment information and frame their curricular planning accordingly (Barnes, Fives, & Dacey, 2015). Such conceptions are best understood as part of an integrated system of individually held implicit or explicit beliefs, which may be subject to change over time and between contexts (Fives & Buehl, 2012). Although a large body of research has been devoted to the study of teachers' conceptions of assessment (e.g. Brown, 2004, 2008; Brown, Hui, Yu, & Kennedy, 2011, 2009; Daniels, Poth, Papile, & Hutchison, 2014; Remesal, 2007; Segers & Tillema, 2011), less is known, however, about their conceptions in relation to standardised tests. While standardised testing has long played a key role in improvement and accountability processes in later grades, several authors note that it has gradually taken a more important role in early childhood (EC) education in the U.S. (Bassok, Latham, & Rorem, 2016; Meisels, Steele, & Quinn, 1989), England (Roberts-Holmes & Bradbury, 2016) and Australia (Kilderry, 2015). Similarly, standardised tests have become widely used in Dutch preschool

CONTACT Niek Frans N.Frans@rug.nl Special Needs Education and Youth Care, University of Groningen, Grote Rozenstraat 38, Groningen 9712 TJ, Netherlands

and kindergarten classrooms (Gelderblom, Schildkamp, Pieters, & Ehren, 2016; Veldhuis & Van den Heuvel-panhuizen, 2014) which is the context of the current study. Two driving factors behind this increasing role are the growing conviction that experiences in EC have a significant impact on later development, along with a trend in educational accountability that has slowly trickled down through primary education (Bordignon & Lam, 2004; DeLuca & Hughes, 2014). Whether these instruments are primarily seen as accountability devices or as efforts to be more responsive to a child's needs could have a substantial influence on the impact that standardised testing has on contemporary EC education (Bassok et al., 2016).

In this article, we define a standardised test as a test that is administered and scored in a methodical manner to produce a score that can be compared to a predefined population (norm-referenced) or some predetermined criterion (criterion-referenced). Although such tests are often described as summative accountability instruments, improvement and accountability purposes are neither mutually exclusive nor inherent to the assessment instrument. As observed by Newton (2007), summative accountability refers to a type of assessment *judgement*, while formative improvement refers to a type of assessment *use*. Given that these two purposes describe 'qualitatively different categories' (Newton, 2007, p.156) norm-referenced or criterion-referenced scores (i.e., summative judgements), which are generally a central aspect of standardised tests, may be employed for formative purposes. As any use of assessment results inevitably includes some form of judgement, a formative use may also be seen as an extension of a summative judgement (Taras, 2005).

While the summative judgement and any additional information that standardised tests provide may be used for improvement purposes, it is crucial to consider whether teachers are able to use instruments that have a clear accountability purpose to serve aims of improvement as well. Brown and Harris (2009) sought to answer this question by studying primary teachers' conceptions of a national norm-referenced adaptive instrument implemented in New Zealand: the Assessment Tools for Teaching and Learning (asTTle). The results indicated that, even though the instrument was designed with the explicit focus on assessment aimed at improvement in learning and teaching, teachers still regarded the instrument as having the primary purpose of 'holding schools accountable.' Coincidentally, the asTTle was utilised predominantly for reporting school quality. Further interviews with teachers, mindfully selected on their questionnaire responses, revealed that some teachers experienced the purpose of demonstrating school competence and quality in a negative way that was contradictory to the use of the same results for improvement purposes. Other teachers, however, did not experience this conflict between two purposes within the same instrument, regarding it instead as a legitimate means of improving instruction and demonstrating accountability (Brown & Harris, 2009). Although both groups of teachers held the conception that the main purpose of assessment was 'to hold schools accountable,' they differed notably in how they experienced this purpose in the asTTle. This outcome demonstrates that teachers' conceptions of assessment have an important affective component in (Fives & Buehl, 2012).

Based on their findings, Brown and Harris (2009) conclude that the assessment format has an impact on assessment use and teachers' conceptions. The formal test-like nature of the asTTle is primarily associated with accountability, while other more informal assessment practices (e.g., observation) are linked to improvement. However, findings on teachers' conceptions of assessment have proven to be highly sensitive to contextual differences (e.g. Barnes, Fives, & Dacey, 2017; Bonner, 2016; Daniels et al., 2014). For

example, Brown et al. (2011) report that teachers in China strongly associated improvement purposes with formal accountability assessment. Conversely, this association was far weaker in the low-stakes context of New Zealand. Differences in teachers' conceptions have also been related to differences in grade level (Bonner, 2016; Brown, 2008). According to Bonner, higher grade levels are generally more accountability-orientated than lower grade levels are. These findings stress the important role of contextual differences in assessment policy and grade in teachers' conceptions of assessment.

Building on the findings reported by Brown and Harris (2009), this study explores EC educators' conceptions of standardised norm-referenced testing in an EC setting. Although the findings reported by Brown and Harris indicate that the majority of teachers viewed 'holding schools accountable' as the primary purpose of these instruments, their conceptions generally differed according to educational stage (Bonner, 2016). It is interesting to see how teachers view such instruments in contexts were assessment for accountability purposes is traditionally less prominent. In addition, while Brown and Harris showed that similar conceptions about the purpose of assessment can be experienced in a highly diverse way, the individual and contextual factors that influence these experiences remain unclear (Bonner, 2016; Brown & Harris, 2009). This study investigates the following two research questions: 1) To what degree do EC educators view a norm-referenced test as an instrument that can serve the purposes of improvement and/or accountability? 2) Which aspects play a role in the differing experiences that teachers have of standardised (norm-referenced) testing? A mixed-method approach was used to build a conceptual framework about the internal and contextual reasons that play a role in teachers' experiences of these instruments. Previous studies have demonstrated the importance of educational context in the study of teachers' conceptions. We therefore start by describing the context of EC education in the Netherlands, as well as its assessment climate.

Study context

In the Dutch system, formal education is compulsory starting at five years of age, although almost all children (99.6%; European Commission/EACEA/Eurydice, 2015) start formal education at four years of age. Since 1985, the two years preceding primary education (ages 4–6; preschool/kindergarten) take place in a school setting [*basisonderwijs*] in which a holistic approach to education has been adopted to support the cognitive, social, and emotional development of children (Dutch Eurydice Unit, 2007). More formalised primary education (ISCED 1) starts around six years of age, when students enter first grade. Assessment in preschool/kindergarten [*kleuteronderwijs*] consists primarily of teacher observation (Dutch Eurydice Unit, 2007). Until 2013, at least one nationally norm-referenced assessment for both language and mathematics was mandated before first grade. Although this directive was changed in 2013, many preschool/kindergarten teachers (>80%) continue to administer nationally norm-referenced tests from the Student Monitoring System [*Leerling- en OnderwijsVolgSysteem*, LOVS] developed by Cito (Gelderblom et al., 2016; Veldhuis & Van den Heuvel-panhuizen, 2014).

The preschool/kindergarten tests of the LOVS are norm-referenced standardised multiple-choice tests. They are typically administered biannually by the classroom teacher, either individually on a computer or using paper-and-pencil forms in a group. The preschool/kindergarten language instruments (Lansink & Hemker, 2012) measure

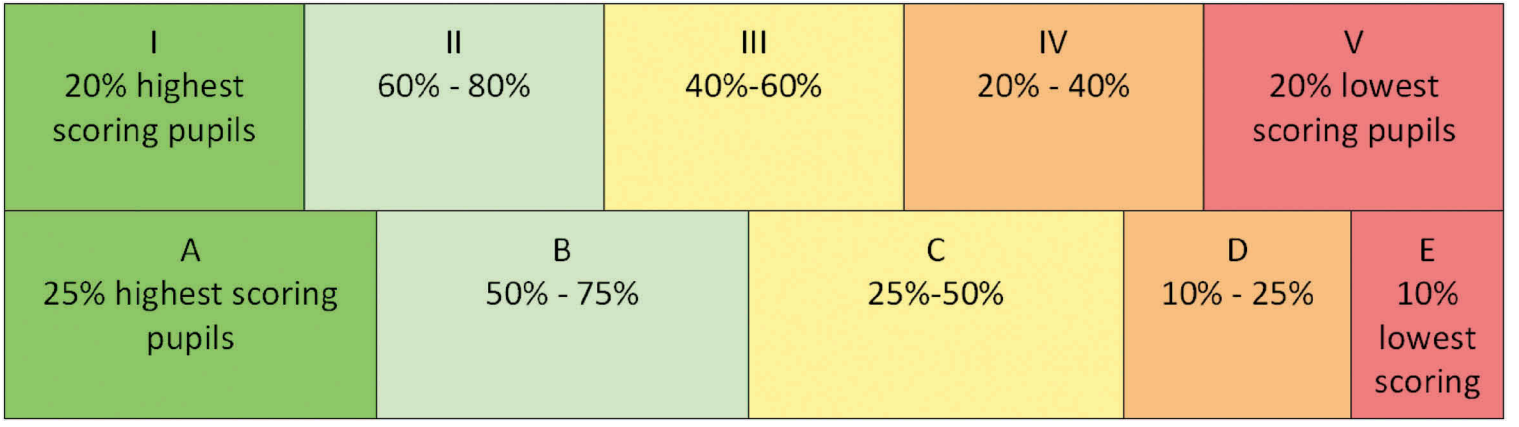

Figure 1. Achievement levels for preschool/kindergarten tests according to the old (bottom) and new (top) distributions, as depicted in the Cito LOVS. Colours vary according to the software used.

receptive language ability and assess the child's performance on six categories: receptive vocabulary, comprehension of spoken language, sound and rhyme, recognition of first and last words, phonemical synthesis, and knowledge of written text. Tasks in the last four categories appear only in the kindergarten test. The mathematics tests (Koerhuis & Keuning, 2011) are designed to measure general emerging numeracy, assessing the child's performance on three categories: number sense, measurement, and geometry.

The official goal of these instruments is two-fold: scores can be used to determine the child's language or mathematics ability as well as the child's progress over time between preschool and kindergarten (Koerhuis & Keuning, 2011; Lansink & Hemker, 2012). A third reported goal that lacks scientific support, is determining areas of over- or underperformance relative to a child's overall ability. The tests are calibrated on large representative samples using Item Response Theory (IRT) to allow comparison of a child's ability and progress to national standards. To facilitate interpretation, the standardised scores are transformed into five achievement levels, ranging from I to V (new classification, since 2013) or from A to E (old classification) as shown in Figure 1. Finally, sub-scores for each category within the test indicate relative strengths or gaps in performance. The test results of children who show low performance or progress can be studied using this 'category analysis' to indicate starting points for intervention (Vlug, 1997). Scores can be aggregated to the group level to create an overview for an entire class or to make comparisons across grade levels and cohorts. An international description of the entire student monitoring system can be found in Vlug (1997).

Like the aTTle studied by Brown and Harris (2009) these are large-scale standardised instruments that measure academic performance in language and mathematics. While the focus in the design and promotion of both tests is on improvement, it is possible to demonstrate accountability through the referencing of scores to national norms. One major difference is that the Cito preschool/kindergarten tests are administered in an EC context were historically accountability testing has not played a major role. Given the importance of contextual differences to teachers' conceptions of assessment (e.g. Daniels et al., 2014) our study asks how educators in this context view this instrument. Moreover, we explore what aspects play a role in their experience of standardised (norm-referenced) testing to extend current theory about teachers' conceptions of assessment.

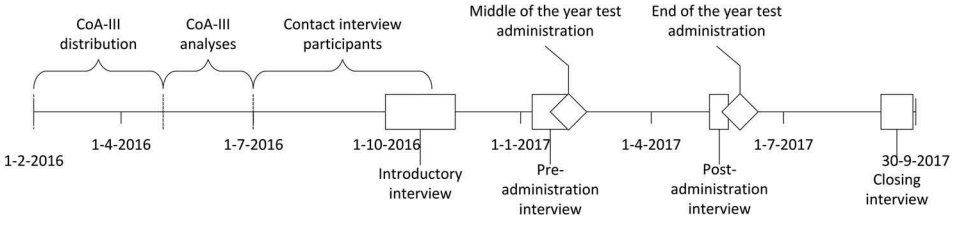

Figure 2. Timeline of the study procedure, date format dd-mm-yyyy.

Method

Population and sample

The sample was recruited from schools that had participated in an earlier study on the stability of the preschool/kindergarten tests in the student monitoring system (Frans, Post, Oenema-Mostert, & Minnaert, 2018) and consisted of 97 participants. Sixty-three percent of the participants were preschool and/or kindergarten teachers, 30% care coordinators, and 3% combined the two functions in part-time appointments. Nearly all of the participants (99%) were women, as is typical of preschool and kindergarten teachers. The age of the participants ranged from 24 years to 64 years, with a median of 50 years. The teachers' experience in preschool and/or kindergarten ranged from two to 45 years, right skewed with a median of 17.5 years. A purposive sample of teachers with varying conceptions of assessment was selected for further interviews. This technique was chosen to capture the entire range of perspectives related to these tests among preschool and kindergarten teachers. The selection was limited to teachers who had provided their email addresses (n= 36). Four teachers either did not respond (n= 1) or declined to participate due to the time investment (n= 2) or retirement (n= 1). The selection procedure is described further in the Results section.

Instrument and design

Conceptions of standardised testing were measured using the Conceptions of Assessment Abridged questionnaire (CoA-III-A; Brown, 2006). Widely used in previous studies, this instrument measures teachers' conceptions about four purposes of assessment: assessment holds schools accountable, assessment holds students accountable, assessment informs the improvement of education, and assessment is irrelevant. Participants were explicitly instructed to answer the statements with the preschool/kindergarten tests designed by Cito in mind, in order to address conceptions of this specific instrument.

The semi-structured interviews were conducted with a subsample of participants over the course of one school year. Because multiple interviews were conducted, the interviewer had more time to build rapport with the participants, and both parties had the opportunity to revisit and further explore topics from the previous interview. Figure 2 presents a timeline of the study procedure. In the introductory interview, teachers were asked to elaborate on their questionnaire answers and experiences with the preschool/kindergarten Cito tests. Each subsequent interview started with the general question of whether the teachers would like to expand on topics from the previous interview or if anything had happened that was relevant to their opinions. Next, teachers were asked to elaborate on any answers that they had given in the previous interview that were unclear or incomplete. The pre-administration interview focused on teachers' conceptions about the test administration and how they perceived the main function of the test for themselves and others. In the post-administration interview, teachers were asked about their experiences with administering the tests, as well as with the results and any subsequent actions that had been taken. The closing interview was used to discuss statements of other teachers that either contrasted with their own or that had not come up in previous interviews. Overall, 24 hours of audio data were collected. Interviews lasted between 34 and 80 minutes, with an average of 60 minutes per interview. Field notes were kept during and directly after each interview.

Procedure

An online version of the questionnaire was sent to a contact person (usually the school director) with the request to distribute the questionnaire to the special services coordinators and preschool and kindergarten teachers in their school. Participants were not informed about any conclusions from the previous study that could influence their responses to the questionnaire. Data on the participants' gender, age, and position were collected, as well as their number of years of teaching experience in preschool/kindergarten. Informed consent was obtained before the start of the questionnaire, and each participant was asked to enter an email address for further contact.

After analysing the questionnaires, teachers with varying conceptions were contacted for participation in the interviews. Participation was voluntary, and no specific details about their questionnaire responses were given until the end of the last interview. Written informed consent was obtained prior to the first interview. The first author conducted the interviews close to the test administrations, so that it would be easier for teachers to relate the interviews to their experiences with the test. All interviews were recorded and transcribed verbatim by an undergraduate student. Each transcript was then compared to the corresponding audio files by the first author and revised as necessary. The revised transcripts were sent to the participants to allow them to correct or reformulate answers in the next interview. Transcripts and field notes were reviewed prior to each interview. Member checks occurred verbally after the last interview, as well as by sending a version of the final report to each participant. Participants were debriefed after the last interview.

Analyses

Confirmatory and exploratory Mokken scale analyses were used to examine the scalability of items and participants on the subscales defined by Brown. The Mokken IRT model, executed with the mokken package in R (Van der Ark, 2007, 2012), permits an assessment of the dimensionality of the data, in addition to providing a means of ordering participants and items simultaneously on each dimension. The model assumes that the probability of endorsing an item is dependent on the degree of a participant's latent trait. The more of the latent trait a person has, the higher the chance of endorsing an item (monotonicity). When items form a perfect Guttman scale, participants who respond negatively to agreeable items will respond negatively to items on the same scale that are less agreeable. This is indicated by a scalability coefficient (H) that typically ranges from zero (no correlation) to one (perfect Guttman scale). The explorative analysis uses the AISP algorithm described by Sijtsma and Molenaar (2002, pp. 71–72). This is a bottom-up procedure that starts by selecting a pair of items that has the highest H coefficient and continues until no items can be found that satisfy a H coefficient higher than a chosen lowerbound c. Values for c between 0 and .55 at increments of .05 were chosen to assess the dimensionality of the data (Sijtsma & Molenaar, 2002). Because the items on the *irrelevance* subscale were negatively worded, the coding of these items was reversed. Interview participants were selected based on the rank orders of their sum scores on the resulting Mokken scales, with the aim of creating maximum variation in the conceptions of interview participants.

The first round of interviews was open coded independently by the first author and an undergraduate student of educational sciences, who transcribed all of the interviews and was trained in qualitative research and the topic of EC education. All sentences pertaining to the preschool/kindergarten Cito test were coded in ATLAS.ti 8. Each quotation received a unique identification number that refers to the interview number (1 to 24) and the quotation number within that interview. A colon separates these values. The first two rounds of interviews were coded in an iterative process, with each interview coded independently. After the codes were discussed and revised, the updated coding scheme was then used in the next interview, and the cycle was repeated. After coding the second interview round, the codes were reorganised by independently clustering related codes and comparing and discussing both schemes. In addition, field notes and memos were reviewed and used to guide this process. In this manner, clusters were formed both inductively from the codes and deductively from field notes and memos kept by the first author. The resulting clusters were discussed among the authors while coding the third and fourth interviews. In order to develop a better idea of relationships between the various themes, paragraphs were coded instead of sentences. Once the coding scheme was complete, the initial interviews were reviewed according to the updated coding scheme. Given that each participant was interviewed repeatedly, it was assumed that the teachers would reproduce important codes and connections. As such, co-occurrences of codes were inspected over all interviews, as well as separately for each teacher, starting with the general themes and ending with individual codes. Prominent co-occurrences were inspected by reviewing the quotations. A conceptual framework was formed in this manner, and other themes that were important to individual teachers were related to this framework.

Results

Participant selection

Analyses of the questionnaire revealed that the irrelevance and student accounting subscales were relatively weak (H= .30 and H= .18 respectively). While the improvement (H= .43) and school accounting (H = .63) scales were stronger, both scales showed a high correlation (r> .60). An exploratory analysis revealed two distinct scales (Appendix A). The first scale (*Relevance:* n= 5, α = .68, H = .34) contains items describing what teachers do and should do with the Cito preschool/kindergarten instruments, and expresses the within-classroom utility of the test. Items on the second scale (*Informative:* n= 24, α = .93,

Table 1. General information on interview participants, percentiles are indicated by P_i.

	Ria	Rianne	Ina	Irina	Renee	Mona
Relevance	P_{81}	P_{76}	P_2	P_{15}	P_{81}	P_{37}
Informative	P_{96}	P_{90}	P_4	P_{63}	P_{24}	P_{51}
Age [years*]	55	25	30	55	45	55
Experience [years*]	25	5	5	25	20	10
Grade level	Kindergarten	Kindergarten	Preschool/Kindergarten	Preschool	Kindergarten	Preschool
Class size*	20	15	10/10	15	15	20
School size*	300 (P_{75})	300 (P_{75})	200 (P_{50})	250 (P_{65})	650 (P_{95})	300 (P_{75})
foreign background*	5% (P_{50})	0% (P_{15})	30% (P_{85})	0% (P_{15})	5% (P_{50})	5% (P_{50})
low educated parents*	5% (P_{45})	10% (P_{70})	15% (P_{80})	5% (P_{45})	5% (P_{45})	5% (P_{45})
Exit score*	P_{65}	P_{70}	P_{10}	P_{45}	P_{75}	P_{65}

Note: Pseudonyms are used for the respondents, numbers in rows with * are rounded to the nearest 5 (50 for school size), in order to preserve confidentiality.

Table 2. Main coding themes related to the preschool/kindergarten Cito tests.

Coding theme	Example
Necessary conditions for testing	'He can do well, if he concentrates.'
Strategies to accommodate conditions	'His mind is somewhere else. Now I've moved him closer to me.'
Target group for test administration	'The children who drop out [score IV/V], they take it again.'
Emotionally charged statement	'Yeah, it's an awful test.'
Relationship to the curriculum	'Not natural to kindergartners. Sitting at a table with a pencil.'
Information gained from the test	'He did better than I thought, because he got a I.'
Alternative means to the test	'But I also use the KIJK and what I observe on my own.'
Professional autonomy of teachers	'We're professional enough to see whether the child can do it or not.'
Purpose according to the teacher	'[Children] who score below average don't meet the standards.'
Expectations of other stakeholders	'They expect group plans to be organised according to the Cito test.'
Use or impact of the test	'You place the weakest in a group, and verify what needs to be practiced.'
Characteristics of the test	'We [test] digitally, but children swipe, while this requires mouse-control.'
Societal context (of the child)	'Every child comes to us differently, some parents don't offer anything.'

H = .40) describe what the test is or does, and portray the degree to which the test results are informative in general.

After analysis of the questionnaire, interview participants were selected to create maximum variation between participant perspectives on both scales. Participant percentile scores for the two scales are presented in the top two rows of Table 1. These scores indicates the percentage of participants with lower scores on the CoA-III-A. For example, Ina's score on the relevance subscale indicates that 2% of the participants ranked lower than her score. Other information on the participants is included to benefit transferability of the results. Besides varied conceptions of the preschool/kindergarten tests, participants varied considerably in terms of age, experience, and grade level taught.

To relate the schools of the participants to the general population in the Netherlands, a comparison was made between the school demographics of the participants' schools and all Dutch primary schools using public databases (DUO, 2017, 2018; RTL, 2017). All six participants teach in Christian schools, which comprise around 60% of all primary schools in the Netherlands. Schools ranged in size from an average number of students (N = 200, Ina) to large schools of 650 students (Renee). Conversely, class sizes vary between 15 and 20 students, which is slightly below the national average of 23 students. It is worth noting that Ina teaches in a mixed classroom of preschool (n= 10) and kindergarten (n= 10) children. With respect to parent education and children with a foreign background, the school population in the schools of Ria, Renee and Mona is representative for the average school in the Netherlands. The schools of Rianne and Irina contain relatively few children with a foreign background, while Ina's school has a large proportion of children with a foreign background. Both the schools of Ina and Rianne contain relatively many children from a low-educated household.

Interviews

Coding of the interviews resulted in 13 themes, presented in Table 2. These themes provide an overview of the entire coding scheme that is included in appendix B. To supplement the questionnaire, different purposes and uses of the Cito preschool/kindergarten test were coded. The following six purposes were mentioned separately by all teachers: The test is 1) a confirmation of the teacher's own judgement; 2) an evaluation of a child's understanding, skill, or ability; 3) a guideline for what a child is expected to learn; 4) a guideline for what

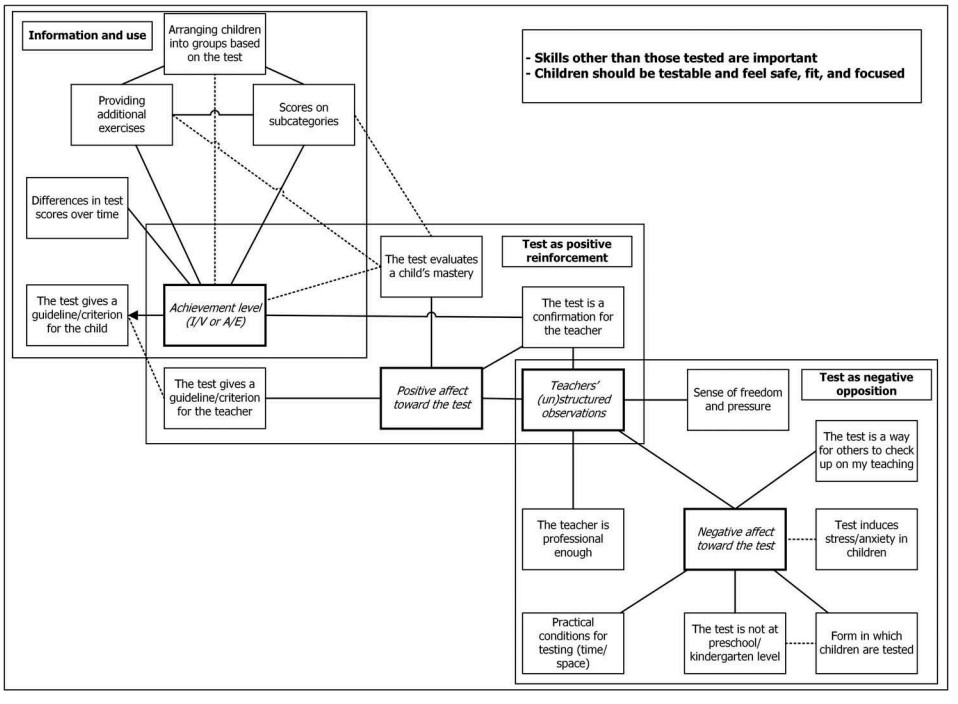

Figure 3. Graphic representation of the main co-occurrences and clusters. Each box represents a separate code. Solid lines indicate co-occurrences reported by at least two teachers. Dashed lines indicate connections that occurred frequently, but not for a particular teacher.

a teacher is expected to teach; 5) an element to consider in decisions concerning skipping or repeating grades; 6) an evaluation of the teacher's ability to teach. Two additional purposes were mentioned by two teachers. Ria regarded the test as pleasant confirmation for the child, and Irina referred to the purpose of familiarising children with formal testing. Reported use of the results was coded separately to distinguish it from potential purposes. All of the teachers reported that they discussed the results with parents and colleagues, in addition to grouping children by achievement level and providing additional exercises. Rianne,

Mona, Renee, and Irina reported at least one instance in which the test was used in decisions to retain or promote children. Ina and Renee reported instances in which test scores were used in decisions to refer children to special education. With the exception of Ria, all of the teachers report adapting the curriculum following a test result.

Although most teachers reported similar conceptions about possible purposes and uses of the test, they differed in their affective reactions. While the themes provide a descriptive account of the topics that the teachers mentioned, they do not depict the relationship between the topics and the teachers' experiences. Exploring the co-occurrence of the individual codes reveals three clusters of codes, as depicted in Figure 3. While all teachers felt that their own observations were dominant in any decision, the perceived relationship between normative scores (achievement levels) and their own observations was a key factor determining how they experienced the tests.

The test as negative opposition to the teacher's own observation

Expressions in this cluster are characterised by a teacher's negative evaluation of the test. Two main conceptions can be distinguished within this cluster. The first concerns the notion that others will use the test as a means of double-checking the teacher's work, such that the test could exert pressure on a teacher's professional identity and sense of freedom in teaching. A second conception in this cluster is that the test is not suitable for young children, either because the form in which children are tested (e.g., 2D paper, multiple choice) feels disconnected from the daily curricular activity or because the test is perceived as being too difficult. This conception is paired with discerned stress and anxiety in children. Finally, practical conditions (e.g., time, functional material, and a suitable room for testing) are often mentioned in this cluster.

> Those categories [achievement levels ed.] for group plans – they don't count for educational inspections, because they don't look at the kindergarten classes (. . .), so I wonder, 'Who are we doing it for?' For our own bit of uncertainty? For the parents who want to see a report? Even though we can create a really nice report with KIJK [structural observation instrument] (. . .) Because, you know, we're not doing the children any favours with that. (Ina, 10:29)

The test as positive reinforcement of the teacher's own observations

This cluster is characterised by a positive experience of the test. The achievement level is conceived primarily as positive confirmation of the teacher's own observations. These ideas are associated with the test as an evaluation of the child's mastery of language and/or mathematics. In this cluster, the test is seen as a positive addition to a teacher's own mental image of a child. This confirmation is closely associated with conceptions of the test as a guideline for the teacher.

> Those tests are fine for checking whether what they've learned is right, (. . .) something like, I expect that this child can actually do really well and always participates well in class – no peculiarities. The child will have a high score. If I can see this result, it provides me with confirmation as a teacher. This is also how it's viewed here at school. (Rianne, 3:7)

Use of the achievement level as a guideline for learning (and teaching)

Most quotations in this cluster concern test use and information in the test. Achievement level plays a central role in test use, and it is often seen as a guideline or criterion for student learning. Test use is mainly described as arranging children into groups according to achievement levels and/or scores on a specific subcategory. Children at low achievement levels are provided additional instruction in small groups. In some cases, changes in achievement between the mid-year administration and end-of-year administration is mentioned as relevant information.

> Right. The small group and then the group table. And those are the children with unsatisfactory scores. They come there and receive additional assignments in the parts on which they did not score well. (Irina, 17:25)

Teacher specific context

The co-occurrences of the codes provide a general framework in which to place teachers' conceptions, including their positive and negative experiences of standardised testing. This demonstrates the central role of the relationship between the teachers' own observations and the normative scores in determining how they experienced the test. Consideration of the specific situations of each teacher made it possible to explore aspects that could help explain why some teachers experienced the test as positive confirmation while others saw it as a negative rejection of their own observations.

In line with her questionnaire responses, Ina's conceptions were wholly contained within the negative-opposition cluster. Ina described how she taught children for whom the highest scores are generally unrealistic expectations. Although she saw considerable development in these children, she did not perceive the test scores as fair reflections of their progress.

> At that time, I had 14 nationalities in my preschool/kindergarten class (…) and they picked up the Dutch language at lightning speed. It was really great to see the strides that they made, and then came that Cito. All of my results were D's and E's (…) not an A anywhere. And that was also a particular community where not everyone wants to live. And then I have to wonder about the standard for this school. It's quite different. (Ina, 4:28)

Given the context in which she was teaching, Ina rarely experienced the test as positive confirmation, instead seeing it as a struggle to keep 'children out of red zone' (Ina, 10:25), referring to the red colour that is used in the computer system to identify the lowest achievement level. The situation was different for Ria, whose classroom scored well above average – a result that she attributes in part to her own enthusiasm as a teacher and that strengthens her image of having a good, well-motivated class.

> I had a question, if you see something like that … with a predominance of green [A level, ed.], what is your conclusion? [Interviewer: You apparently have a high score in the class. That's what I would say. What would your own conclusion be?] Yes. I think so too. (…) This is simply a good class that's motivated (…) and if the teacher is enthusiastic as well … I just have to put two and two together. (Ria, 20:5)

Ina further described how her previous director required the use of the test, while her new director urged teachers to decide for themselves when to administer the test. This change in management considerably influenced her experiences with the test and contributed to her more positive stance in the final interview.

> He does allow us space to brainstorm and think about it, and the previous school manager had imposed it a bit more (…) when it's imposed (…) that incites resistance (…) the sense of confidence in us, that, as teachers, we're professional enough to act and decide, while knowing that these instruments are available and that we can use them. But they're not required. (…) Right. And that feels good, because the teacher has more control, and that's really nice. (Ina, 22:24)

Aside from the compulsory use of the preschool/kindergarten Cito tests, Ina experienced little positive support or interest in the results from her previous management team, which contributed to her initial negative appreciation of the test.

> The previous special services coordinator was a big fan of growth curves, and therefore just wanted to see development of a certain number of points. Well, we did what we were asked

> to do. Thereafter, we weren't asked about it very often, and if we didn't achieve it, that was that. It made me wonder why we were doing it. (…) It's such a shame that the Cito is given to preschoolers/kindergartners – we're doing something with it, because we have to, and so forth. (…) But it's not so binding. It's not anything decisive. (Ina, 4:31)

Although Mona's response to the questionnaire was more neutral than Ina's was, she shared many of the same negative associations. Similar to Ina, her concerns related to the population in her classroom. In Mona's case, however, it had to do with the age of the children in her class. Both Mona and Ina reported the negative experience that the test induced stress in children.

> I'm happy that we did not do the tests, because it was also very frustrating for the little ones. They came in, and they had to sit down at a little desk, with a sheet of paper in front of them (…) they do have to do that in kindergarten, but then they're already somewhat more advanced in their development. They are (…) more ready than they were in preschool. (Mona, 7:7)

Unlike Ina, Mona has complete freedom with regard to when and whom to test. Although she reported having negative experiences with mandatory classroom-wide administration in the past, she now saw the tests as contributing positively to her own observations in case of uncertainty concerning a child's abilities.

> Right. I don't think we can let go of it. I think that we can't just have an idea in mind, but want some confirmation – through that test. It's sometimes really nice to know (…) but do it for the children we're not sure about. (Mona, 7:24)

Ina and Mona shared the conception that the test was unsuitable for children in preschool, and they both saw the tests as more appropriate for children in kindergarten.

> I think it's a good thing in kindergarten, because they'll be going to first grade, and then they will be expected to know a thing or two, as the Cito tests will be used from then on. (…) Right. That's a condition, and they pay a lot more attention to it in kindergarten than they do in preschool. (Mona, 7:4)

> I think it's a lot more useful in kindergarten, because preschool (…) I'm happy we don't have it anymore. (…)I also started with it in kindergarten at one time. Then I was able to see the utility of Cito, and now, from this perspective, I am better able to see the utility of Cito. (Ina, 22:1)

In contrast to Ina and Mona, and in accordance with their questionnaire responses, the conceptions of Ria and Rianne are mostly described by the positive reinforcement and use clusters. Both regarded the test as positive confirmation of their teaching, and both experienced it as having a positive effect on children. For Ria, the test functioned as a guideline for what should be taught and learned, in addition to serving as a subsequent evaluation of her performance as a teacher.

> We have to get it right. (…) I'd hate to see children from the school where I'm working be at a disadvantage immediately upon entering secondary school. (…) and the Cito is very strongly oriented to, 'This is the standard, and you have to meet it.' (Ria, 2:16)

She reported believing that a head start in language and maths skills are of vital importance to a child's future education and feeling strongly about using the test as an aim and guideline in her teaching of these skills.

> 'I consider the way it's done much too scholastic' [reading a quote] (…) I don't think that at all. (…) You know. It's not scholastic at all. We start preparing them for something that's

> really difficult. (…) I see it as a challenge. This is where we're heading, isn't it? Isn't it great to go to first grade and learn how to write? (Ria, 20:22)

For Rianne, the test's function as confirmation of her observations was more prominent, although she also perceived the test to serve an evaluative function with regard to her teaching. In contrast to Ina, Rianne described how her Management Team (MT) showed interest in the results and allowed her to take the lead in finding problems and possible solutions.

> They [the special services coordinator and the director] also want to know what we're going to do with it. (…) So, if there are weak students, they want us to tell them what we're planning to do. (…) And then we have to give it a lot of consideration. That's a good thing, though, because we can look at it again and see if we've achieved what we expected. (Rianne, 3:26)

Renee and Irina reported mixed negative and positive experiences with the test. While Renee generally agreed with the test's function as a guideline for her as a teacher, she expressed concerns about the suitability of the form in which children are tested.

> After all, we do have to have a certain standard that we have to meet. (…) But it's obviously all about how we approach it ourselves as teachers (…) Right. It's something to work toward. (…) Well, if all of those children achieve a score of C, then I'd think it would meet the standard. It doesn't have to be all B's and A's, but if we have C's … okay. (…) Yeah, it's more of a guideline. I do think it's important to work in a targeted manner. (Renee, 24:15)

> Well, I don't think it's suitable for preschoolers. (…) it's fine to look at children and say that, at some point, they will have to … (…) But I just don't think the form is right. And for the parents, we obviously have to, we obviously have to … well, have something to show. (…) work more with materials or something like that, be more at the preschool level. I consider the way it's done much too scholastic. (Renee, 12:22)

Similar to Ina, Renee described how her experience with the test depended on her classroom population. While she noted that, in her previous school, she had felt coerced to keep children out of the low scoring categories, she reported noticing a much more positive attitude in her new school, where higher scores were more common.

> Because, in another school (…) there were a lot of ethnic minority children. And then it was important to train them in that, because (…) this one only made a D, but will nevertheless have to go to first grade. Then I tend to think, 'Just forget it. It'll turn out okay.' But at that school, we had to do a lot more … I'm now in a much better social environment, and so it's just much less of an issue. It's obvious that we're much more relaxed about it. But that's because it's possible. Because those children, they'll get there. (Renee, 6:32)

In contrast to other teachers, Irina's conception of the main purpose of the test was unrelated to the child's achievement level. Instead, she reported that the main purpose of the test was to familiarise children with formal testing situations.

> We actually see it more as getting used to (…) taking the Cito. Personally, I don't assign much of a value judgment, like 'Whoa. That's a problem,' or 'That's bad.' I do consider it in my class plan, though. (Irina, 5:3)

Like Mona, she enjoyed considerable freedom in deciding whom to test, as well as in deciding the manner in which she conducts testing. Although the scores of Irina's

classroom were most similar to those of Ina's, she did not report having the same negative experience with the test.

Discussion

One of the goals of this study was to explore the extent to which EC educators view a norm-referenced test as an instrument of improvement and/or accountability. The analyses of the CoA-III questionnaire (Brown, 2006) revealed no clear distinction between the teachers' conceptions of improvement and accountability purposes. The high correlations between the concepts (r~ .60) suggest that educators in this sample either believe the instruments are suitable for both purposes or neither. Findings from the exploratory Mokken scale analysis did suggest that educators made a distinction between the validity and suitability of the information in the test and its usefulness to them. This finding could indicate that educators in this sample share the view of Taras (2005) that judgement and use are complementary parts of the same assessment process. Analysis of the interviews showed that teachers generally identified the tests as serving the same purposes, and they reported using the results in a similar manner. These results suggest that, although teachers are aware of both the accountability and the improvement purposes of the tests, they differ substantially in how they experience and cope with these purposes.

The conceptual framework emerging from the interviews identified the perceived relationship between the test standard and the teacher's own observations (whether structured or unstructured) as a central aspect in teachers' experiences with the preschool/kindergarten tests. While some teachers experienced the normative scores of the test as pleasant confirmation, others experienced them as negative opposition to their own observations. Several aspects seemed to influence how teachers viewed the relationship between their own observations and the test. First, the type of classroom can influence teachers' perceptions of the normative scores. Some teachers (e.g., Ina) never experience the normative score as pleasant confirmation, as the children in their classrooms do not generally score at the higher achievement levels. Even when children show considerable development, a below-average score may feel like a rebuttal of the progress observed by the teacher. When the majority of children in a class score well above average (as was the case for Ria's classroom), teachers are more likely to experience testing as the attainment of success rather than as the avoidance of perceived failure. This finding is congruent with the observation by Harris and Brown (2009), who argue that tests may be perceived as unfair in schools with scores in the lower decile. Specific features of the test (e.g., the colour system for the various achievement levels) may further reinforce the idea that scoring below average (red) is inherently bad, while scoring above average (green) is a goal worth pursuing.

Although Irina did not perceive the test results as being particularly relevant to her teaching, and although her classroom's level of achievement was at a low level similar to that of Ina's classroom, she did not share the same negative experience. Her conception that the primary purpose of the instrument was 'to familiarise children with formal testing situations' might have helped her experience the test in a positive manner. This purpose is fulfilled when children are placed in the testing situation regardless of the results achieved, thereby diminishing the importance of the norm-referenced score to her experience. In

addition, this particular conception meant that she did not experience the test format as in any way unsuitable for young children. Another factor that may have played a role for Irina and other teachers is the support that they received from the school's MT. This was clearly reflected in the interviews with Ina, for whom testing was obligatory and who had experienced little support and interest in monitoring and mediating the outcomes from her previous MT. This had eliminated her sense of agency and professionalism in the assessment process, which she subsequently regained when her new director included teachers in an open discussion on test usage. This result resonates with the finding by Oosterhoff, Minnaert, Oenema-Mostert, and Goorhuis-Brouwer (2014) that school director plays a key role in the perceived autonomy of teachers.

It is important to note that these results are not an evaluation of the quality of the CoA-III. The small number of participants and items restrict us from making any definitive claims about the dimensionality of this instrument. In addition, the specific context of the study and differences in the chosen analytical method make it difficult to relate these findings to previous studies on the CoA-III. The integration of EC education into primary school in the Netherlands may have contributed to a more curriculum-orientated approach (Den Elt, Van Kuyk, & Meijnen, 1996) compared to other countries. Since a description of EC education in each country goes beyond the scope of this paper, we leave it to the reader to compare the context of this study to their own.

The relatively low H values of the two scales used in this study indicates that these scales provide an approximate ranking of participants. While this gives a reasonable selection-criterion for interview candidates, it limits the utility of these scales for other purposes where more precision is required. Since the scalability coefficients are dependent on the number of items in a scale, a comparison between the full questionnaire and the abridged version may be a valuable addition in further research.

The qualitative design of the study focuses on diversity of conceptions rather than generalisability. As such, the results should be seen as extending current theory of teachers' conceptions of assessment. In accordance with Fives and Buehl (2012) the interview results show that teachers' conception of assessment are integrated in a larger system of beliefs about their role as a teacher and what constitutes appropriate assessment for young children. In addition, the results illustrate the importance of contextual demands (Fives & Buehl, 2012) and internal beliefs about the attainability of these demands. Although some teachers viewed the normative scores as a positive confirmation or guideline, it can become a source of frustration in an underprivileged environment. Invariably, teachers spoke in terms of failure if children did not score at least average. This position creates unrealistic expectations for both teachers and children, and sometimes led to curricular decisions that were based on the test form or content. A child-centred norm may provide teachers with the same impression of confirmation whilst avoiding a sense of unfairness or punishment. Finally, although the inclusion of other stakeholders in the assessment process fell outside the scope of this study, including the experiences of parents, management teams and children could provide important insights into the use of standardised tests in EC education.

Acknowledgments

We would like to thank Silke van der Velde for her valuable contribution in transcribing and coding the interviews. We are also grateful to Arda Oosterhoff for her constructive remarks on the Results section and her helpful reflections on the coding scheme.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Groningen through the Netherlands Organization for Scientific Research (NWO) Graduate Program. The NWO had no influence in the study or the decision to publish.

Notes on contributors

Niek Frans is a PhD candidate at the University of Groningen. His research focuses on issues of stability, predictive value and utility in early childhood standardised assessment. This research is conducted within a project on the administration of national standardised norm-referenced preschool tests within the Netherlands.

W. J. Post is a (bio)statistician and associate professor at the University of Groningen. Her research is directed at validity issues, such as statistical conclusion validity (statistical modelling), construct validity (measurement theory; IRT models), internal validity (missing data), and external validity (generalisation, consequential validity).

C. E. Oenema-Mostert focuses on dynamic and authentic development of young children within their natural context as an associate professor at the University of Groningen and Stenden University of Applied Sciences. Her main occupation is research in the field of the playing, learning and developing child between three and seven years in a home-based or centre-based educational environment.

A. E. M. G. Minnaert is a professor at the University of Groningen. His research activities focus on the interplay of social, motivational, emotional and (meta)cognitive factors related to development, learning and achievement in and out of school. Special attention is paid to children with SEN, to their teachers and to (intervention) programmes targeted at the improvement of (environments conducive for) children with SEN.

ORCID

Niek Frans http://orcid.org/0000-0001-6684-0684
W. J. Post http://orcid.org/0000-0002-8655-5204

References

Barnes, N., Fives, H., & Dacey, C. (2015). Teachers' beliefs about assessment. In H. Fives & M. Gregoire Gill (Eds.), *International handbook of research on teachers' beliefs* (pp. 284–300). New York: Routledge.

Barnes, N., Fives, H., & Dacey, C. M. (2017). U.S. teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, *65*, 107–116.

Bassok, D., Latham, S., & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open*, *1*(4), 1–31.
Bonner, S. M. (2016). Teachers' perceptions about assessment: Competing narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (1st ed., pp. 21–39). New York: Routledge.
Bordignon, C. M., & Lam, T. C. M. (2004). The early assessment conundrum: Lessons from the past, implications for the future. *Psychology in the Schools*, *41*(7), 737–749.
Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318.
Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged instrument. *Psychological Reports*, *99*(1), 166–170.
Brown, G. T. L. (2008). Teachers' conceptions of assessment. In T. L. B. Gavin (Ed.), *Conceptions of assessment : Understanding what assessment means to teachers and students* (pp. 91–118). New York: Nova.
Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of Multidisciplinary Evaluation*, *6*(12), 68–91.
Brown, G. T. L., Hui, S. K. F., Yu, F. W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5–6), 307–320.
Brown, G. T. L., Lake, R., & Matters, G. (2009). Assessment policy and practice effects on New Zealand and Queensland teachers' conceptions of teaching. *Journal of Education for Teaching*, *35* (1), 61–75.
Daniels, L. M., Poth, C., Papile, C., & Hutchison, M. (2014). Validating the conceptions of assessment-III scale in Canadian preservice teachers. *Educational Assessment*, *19*(2), 139–158.
DeLuca, C., & Hughes, S. (2014). Assessment in early primary education: An empirical study of five school contexts. *Journal of Research in Childhood Education*, *28*(4), 441–460.
Den Elt, M. E., Van Kuyk, J. J., & Meijnen, G. W. (1996). Culture and the kindergarten curriculum in the Netherlands. *Early Child Development and Care*, *123*, 15–30.
DUO. (2017). *Leerlingen po zoals geregistreerd in BRON*. Groningen: Dienst Uitvoering Onderwijs.
DUO. (2018). *Adressen van alle schoolvestigingen in het basisonderwijs*. Groningen: Dienst Uitvoering Onderwijs.
Dutch Eurydice Unit. (2007). *The education system in the Netherlands 2007*. Brussels: The Hague.
European Commission/EACEA/Eurydice. (2015). *Early childhood education and care systems in Europe. National information sheets – 2014/15*. Brussels: European Commission. doi:10.2797/48986
Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Individual differences and cultural and contextual factors* (Vol. 2, pp. 471–499). Washington, DC: APA.
Frans, N., Post, W. J., Oenema-Mostert, C. E., & Minnaert, A. E. M. G. (2018). *Defining and evaluating stability in assessment*. Under review.
Gelderblom, G., Schildkamp, K., Pieters, J., & Ehren, M. (2016). Data-based decision making for instructional improvement in primary education. *International Journal of Educational Research*, *80*, 1–14.
Harris, L. R., & Brown, G. T. L. (2009). The complexity of teachers' conceptions of assessment: Tensions between the needs of schools and students. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 365–381.
Kilderry, A. (2015). The intensification of performativity in early childhood education. *Journal of Curriculum Studies*, *47*(5), 634–653.
Koerhuis, I., & Keuning, J. (2011). *Wetenschappelijke verantwoording van de toetsen Rekenen voor kleuters*. Arnhem: Cito.

Lansink, N., & Hemker, B. T. (2012). *Wetenschappelijke Verantwoording van de toetsen Taal voor kleuters voor groep 1 en 2 uit het Cito Volgsysteem primair onderwijs*. Arnhem: Cito. Retrieved from http://toetswijzer.kennisnet.nl/html/tg/18.pdf

Meisels, S. J., Steele, D., & Quinn, K. (1989). *Testing, tracking, and retaining young children: An analysis of research and social policy*. Washington D.C.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170.

Oosterhoff, A., Minnaert, A. E. M. G., Oenema-Mostert, C. E., & Goorhuis-Brouwer, S. (2014). *Constrained or sustained by demands? Personal feelings of professional autonomy in early childhood education. Poster session presented at the 24th EECERA conference*. Crete, Greece.

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *Curriculum Journal*, *18*(1), 27–38.

Roberts-Holmes, G., & Bradbury, A. (2016). The datafication of early years education and its impact upon pedagogy. *Improving Schools*, *19*(2), 119–128.

RTL. (2017). *Eindtoets cijfers*. Hilversum.

Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation*, *37*, 49–54.

Sijtsma, K., & Molenaar, I. W. (2002). *MMSS introduction to nonparametric item response theory*. London: Sage Publications, Inc.

Taras, M. (2005). Assessment – Summative and formative – Some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. Retrieved from: http://www.jstatsoft.org/v20/i11/

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. Retrieved from: http://www.jstatsoft.org/v48/i05/

Veldhuis, M., & Van den Heuvel-panhuizen, M. (2014). Primary school teachers' assessment profiles in mathematics education. *PloS One*, *9*, 1.

Vlug, K. F. M. (1997). Because every pupil counts: The success of the pupil monitoring system in The Netherlands. *Education and Information Technologies*, *2*, 287–306.

Appendix A.

Exploratory Mokken scale analysis with item H coefficients for the two scales.

	Brown (2006) factor structure			
Scale: The Cito test is useful for teachers (H = .34)	Second order	First order	H_i	(*SE*)
Assessment results should be treated cautiously because of measurement error*	Irrelevant	Inaccurate	.38	(.070)
Assessment forces teachers to teach in a way against their beliefs*	Irrelevant	Bad	.37	(.066)
Teachers should take into account the error and imprecision in all assessment*	Irrelevant	Inaccurate	.36	(.068)
Assessment results are filed and ignored*	Irrelevant	Ignored	.30	(.105)
Teachers conduct assessment but make little use of the results*	Irrelevant	Ignored	.29	(.087)
Scale: The Cito test provides valid information (H= .40)				
Assessment provides information on how well teachers are doing	.	.	.50	(.041)
Assessment helps students improve their learning	Improvement	Improves learning	.48	(.047)
Assessment is a way to determine how much students have learned …	Improvement	Describes abilities	.48	(.042)
Assessment is a good way to evaluate a school	School Accounting	-	.48	(.048)
Assessment is an accurate indicator of a school's quality	School Accounting	-	.46	(.053)
Assessment establishes what students have learned	Improvement	Describes abilities	.45	(.055)
Assessment results can be depended on	Improvement	Valid	.45	(.050)
Assessment is an accurate indicator of a teacher's quality	.	.	.45	(.044)
Assessment results are trustworthy	Improvement	Valid	.43	(.052)
Assessment feeds back to students their learning needs	Improvement	Improves learning	.42	(.054)
Assessment results are consistent	Improvement	Valid	.42	(.055)
Assessment is integrated with teaching practice	Improvement	Improves teaching	.42	(.047)
Assessment determines if students meet qualification standards	Student Accounting	-	.42	(.055)
Assessment is a good way to evaluate a teacher	.	.	.42	(.044)
Assessment provides information on how well a group is doing	.	.	.42	(.046)
Assessment is an imprecise process*	Irrelevant	Inaccurate	.41	(.052)
Assessment provides information on how well schools are doing	School Accounting	-	.38	(.058)
Assessment is unfair to students*	Irrelevant	Bad	.35	(.061)
Assessment measures students' higher order thinking skills	Improvement	Describes abilities	.34	(.064)
Assessment information modifies ongoing teaching of students	Improvement	Improves teaching	.33	(.074)
Assessment allows different students to get different instruction	Improvement	Improves teaching	.33	(.070)
Assessment provides feedback to students about their performance	Improvement	Improves learning	.30	(.056)
Assessment interferes with teaching*	Irrelevant	Bad	.30	(.068)
Assessment is assigning a grade or level to student work	Student Accounting	-	.25	(.073)

Note: The items 'Assessment places students into categories' and 'Assessment has little impact on teaching' did not fit any scale. Reverse-coded items are indicated with *. Items added to the questionnaire are indicated by a '.' in the original factor. Removing these items had no effect on the scales.

Appendix B.

Full codebook

Coding theme	Main codes (number of subcodes, if applicable)	Example
Necessary conditions for testing	Conditions for testing related to the child (7) Practical conditions for testing Conditions related to the teacher (2)	The child needs to be able to focus The test takes a lot of time You need to know the manual somewhat
Strategies to accommodate conditions	Before test administration During test administration After test administration	We avoid the word 'test' Children that have trouble concentrating sit close to me If a child is anxious I re-test him or her one-on-one
Target group for test administration	Dependent on the grade-level Dependent on the previous test score Dependent on grade retention Dependent on confidence teacher Dependent on parent request	We don't administer the test in preschool We re-test children in June if they score a D/E in January We don't test a child that is going to repeat kindergarten I only test a child when I have doubts about his/her level I sometimes re-test when the parents ask me to
Emotionally charged statement	Positive affect teacher Negative affect teacher Positive affect child Negative affect child Positive affect other stakeholder Negative affect other stakeholder	I'm glad that we have this test It's a horrible test Children love working in a booklet Children get stressed and anxious when tested Parents think the test is important Some of my colleagues hate these tests
Relationship to the curriculum	Play should be central in the curriculum Cognitive challenge is important in (pre-)K Education should be child directed The test is not on (pre-)K level Other skills than those tested are important	Children learn mainly by playing Challenging children in language is vital at a young age Children will ask about writing when they are ready The level of the test is too difficult for many children His score is good but he still acts too young for his age
Information gained from the test	Achievement level Differences in test scores over time Scores on subcategories Observation during the test administration Answers to specific questions	This child scored a D on language You can see that her achievement score has gone up If numerical understanding is low, you can focus on that I noticed that he is unable to listen to my instructions You can see here that he worked from right to left
Alternative means to the test	(Un)structured observation by the teacher Teacher designed tests Other external tests	You have your own observations which tell you a lot I gave them a small task to see if they could do it We also have a vocabulary test in October

(*Continued*)

(Continued).

Coding theme	Main codes (number of subcodes, if applicable)	Example
Professional autonomy of teachers	Teacher's sense of trust Teacher's sense of professionalism Teacher's sense of freedom and pressure Teacher's sense of autonomy support	Why do you need the test, trust the teacher for once And it's like someone wants to check if I'm good enough I am forced to administer this test Sometimes you miss things that the test helps you see
Purpose according to the teacher	Confirmation for the teacher/child (2) Evaluation of the teacher/curriculum (1) Evaluation of the child's mastery (3) Guideline for what a child is expected to know Guideline for what a teacher is expected to teach Familiarising children with formal testing Indication for grade-skipping or retention	The test is a confirmation for you as a teacher The test can tell me if what I offered was sufficient The test shows what a child can and cannot do The test shows what a child needs to know I look at the test to see what needs to be taught For us the main idea is that children get used to testing I would be hesitant to send him to first grade with two D's
Expectations of other stakeholders	Control/confirmation of educational process Making the grade (scoring at least average) Growth between test administrations Few or no expectations	Parents want to know a child's level, the test provides this The school wants to know if I am on par with expectations Parents want to see if their child has grown The educational inspection is not interested in the results
Use or impact of the test	Use of the results (7) Impact of the test on education (2) Impact of the test on the behaviour of others Limited impact or use of the test	We use the results to arrange children into groups I teach the word 'Antlers' as they often struggle with it Parents practice at home so their child scores higher In practice you don't do a lot with the results
Characteristics of the test	Form in which children are tested Content of the test Administration of the test (2) Continuity between tests	Assignments in the test are all in 2D Questions can be interpreted in multiple ways The test is just a snapshot of the child's development The tests between years are so different
Societal context (of the child)	Background/context of the child Higher external demands	Some children just learn more from their parents Society just expects more nowadays